УДК 004.89:004.93

А.О. Журавлев
Институт проблем искусственного интеллекта
МОН Украины и НАН Украины, г. Донецк
Украина, 83048, г. Донецк, ул. Артема, 118 б

Создание базы данных моделей управления слов русского языка

A.O. Zhuravlyov
Institute of Artificial Intelligence MES of Ukraine and NAS of Ukraine, c. Donetsk
Ukraine, 83048, c. Donetsk, Artema st., 118 b

Creating of Database of Governance Models for Russian Words

А.О. Журавльов
Інститут проблем штучного інтелекту МОН України і НАН України, м. Донецьк
Україна, 83048, м. Донецьк, вул. Артема 118 б

Створення бази даних моделей управління слів російської мови

Статья посвящена извлечению знаний из лингвистических словарей для наполнения базы данных моделей управления слов русского языка. Разработана схема использования базы данных моделей управления слов русского языка при выполнении семантико-синтаксического анализа предложений, выполнено проектирование базы данных моделей управления, разработана методика автоматизированного наполнения базы данных моделей управления слов русского языка.
Ключевые слова: автоматическая обработка текста, модель управления слова, семантико-синтаксический анализ, база данных, лингвистические словари.

The article describes knowledge extracting from language dictionaries to fill the database of governance models for Russian words. It is developed a scheme for using the database by systems of semantic and syntactic analysis, designed database structure, proposed the methods of automated filling of the database of governance models for Russian words.
Key words: automatic text processing, governance model of word, semantic and syntactic analysis, database, linguistic dictionaries.

Статтю присвячено видобуванню знань з лінгвістичних словників для наповнення бази даних моделей управління слів російської мови. Розроблено схему використання цієї бази даних при виконанні семантико-синтаксичного аналізу речень, спроектовано базу даних моделей управління, розроблено методику автоматизованого наповнення бази даних моделей управління слів російської мови.
Ключові слова: автоматична обробка тексту, модель управління слова, семантико-синтаксичний аналіз, база даних, лінгвістичні словники.

Введение

В настоящее время возникает необходимость в разработке программных средств автоматической или автоматизированной обработке естественно-языковых (ЕЯ) текстов русского языка. Например, при сборе и фильтрации данных из различных источников, извлечении знаний, реферировании, аннотировании и т.п. Одним из ключевых этапов обработки ЕЯ текстов является синтаксический анализ.

К настоящему времени опубликовано множество словарей, описывающих лексико-грамматические средства выражения семантико-синтаксических связей в предложении. К их числу относятся толково-комбинаторные словари [1], синтаксические словари [2], семантические словари [3], словари управления [4].

Знания, представленные в этих и подобных им словарях, необходимы при создании лингвистических процессоров и других систем, предполагающих выполнение семантико-синтаксического анализа текстов. В связи с этим извлечение знаний из текстов упомянутых словарей является актуальной задачей, направленной на: развитие методов и средств компьютерной лингвистики, создание прикладных систем автоматической обработки ЕЯ текстов.

Потребность в средствах семантико-синтаксического анализа текста, опирающихся на лингвистические знания, огромна. Основной трудностью на пути их создания является плохая формализованность языка, отсутствие общедоступных лингвистических баз данных и знаний. Одной из важных подсистем модуля семантико-синтаксического анализа, способной повысить эффективность его работы, является база данных моделей управления (МУ) слов русского языка.

Цель работы: разработка методики извлечения знаний из лингвистических словарей для наполнения базы данных моделей управления слов русского языка.

Для достижения поставленной цели необходимо решить следующие задачи: разработать схему использования системы, реализующей базу данных МУ слов русского языка при выполнении семантико-синтаксического анализа предложений русского языка; выполнить проектирование базы данных МУ; разработать методику автоматизированного наполнения базы данных МУ слов русского языка.

Проектирование базы данных моделей управления слов русского языка

МУ слова – одно из важнейших лексикографических понятий. С помощью МУ в комбинаторных словарях пытаются представить одновременно синтаксические и семантические валентности слова. Для большинства предикатных слов число семантических и синтаксических валентностей одинаково и совпадает, соответственно, с числом мест в МУ [1].

На данный момент даже для английского языка не существует «прикладных программ, использующих методы искусственного интеллекта, способных нетривиально перерабатывать извлеченные из текста элементы знаний (интерпретировать, обобщать, выявлять зависимости, прогнозировать и т.п.)» [5]. Такая ситуация обусловлена, по-видимому, следующими причинами.

1. Мало распространены системы лингвистического анализа текста, способные интерпретировать отношения ассоциативной связи между словами, то есть извлекать знания как некоторые элементы, обладающие внутренней структурой и пригодные для нетривиальной смысловой обработки.

2. Алгоритмы семантического анализа текстов слабоэфективны, из-за низкой достоверности автоматически извлекаемых утверждений и фактов, что объясняется несовершенством алгоритмов семантического анализа и низким качеством источников информации.

Технология использования моделей управления для анализа ЕЯ текстов применима в системах семантико-синтаксического анализа текстовых документов.

Предполагаемая схема использования системы, реализующей базу данных МУ слов русского языка при выполнении семантико-синтаксическом анализа предложений, представлена на рис. 1.

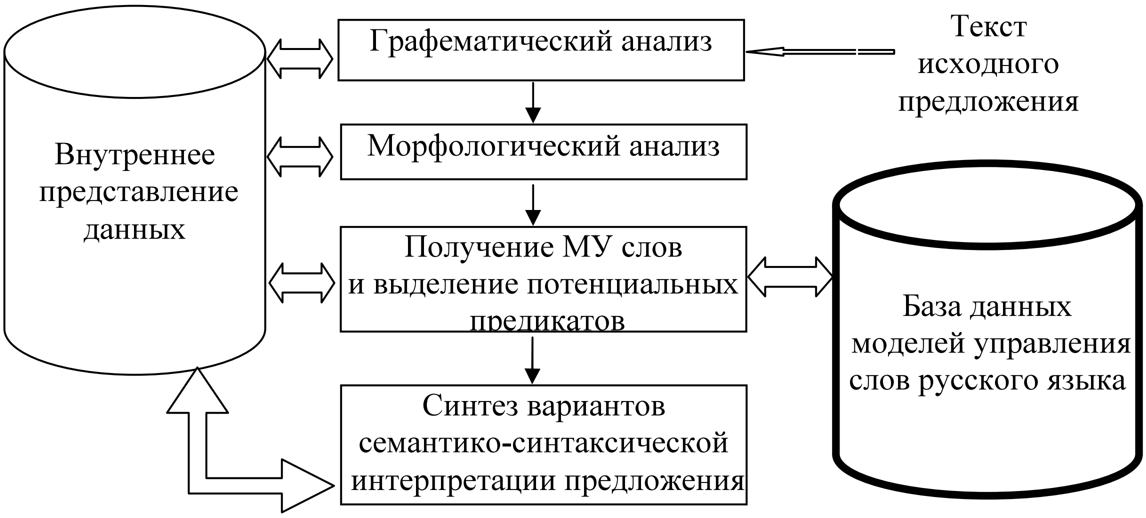


Рисунок 1 – Схема использования базы данных моделей управления слов русского языка при семантико-синтаксического анализе предложений

Анализ предложения начинается с подачи исходного текста в модуль графематического анализа, в задачу которого входят: разделение входного текста на слова, выделение устойчивых оборотов, не имеющих словоизменительных вариантов, выделение дат в цифровых форматах, выделение ФИО (фамилия, имя, отчество), когда имя и отчество написаны инициалами, выделение электронных адресов и т.п.

Полученные данные подаются в модуль морфологического анализа, где для каждого слова определяют множество вариантов интерпретации. Вариант интерпретации представлен в виде пары – написания леммы и морфологической информации словоформы.

Результаты работы модуля морфологического анализа в виде последовательности векторов множеств интерпретаций словоформ, составляющих предложение, подаются на вход модуля получения МУ слов и выделения потенциальных предикатов, который обращается к базе данных МУ слов русского языка.

По полученным вариантам МУ слов и предикатам модуль синтеза вариантов семантико-синтаксической интерпретации предложения генерирует семантико-синтаксические представления предложения в виде деревьев синтаксического подчинения (с учетом явления омонимии вариантов представления может быть несколько).

Исходя из предложенной схемы системы семантико-синтаксического анализа предложений русского языка, для обеспечения ее корректной работы каждая запись в разрабатываемой базе данных МУ должна содержать следующие поля:

– написание слова с указанием ударения;
– морфологическая информация;
– толкование слова;
– семантический класс;
– список МУ.

Ударение и толкование не являются обязательными элементами, их необходимо заполнять для снятия омонимии на морфологическом, синтаксическом и семантическом уровнях анализа предложения.

Каждая МУ представляет собой список валентностных гнезд, для которых указаны падежи и управляющие ими предлоги. В ряде случаев валентностное гнездо может быть заполнено более чем одним актантом. Некоторые из актантов могут быть не обязательными. Для хранения МУ будем использовать таблицу со следующими полями:

– номер модели управления;
– номер валентностного гнезда;
– номер множества предлогов;
– номер падежа.

Такая организация таблицы моделей управления позволяет хранить несколько альтернативных вариантов заполнения одного валентностного гнезда в рамках одной модели управления.

Автоматизированное наполнение базы данных моделей управлений слов русского языка

В качестве источников информации для заполнения базы данных можно использовать толково-комбинаторные, синтаксические, семантические словари, словари управления. В данной работе в качестве источника информации для заполнения базы данных МУ слов русского языка был выбран словарь Розенталя [4], электронная версия корого находится в свободном доступе.

Основу словарной статьи этого словаря составляют:

– заголовочное слово;
– местоименные вопросы к заголовочному слову, по которым можно определить семантическое значение слова, следовательно, при наличии семантической классификации слово можно отнести к определенному семантическому классу;
– иллюстративные примеры, которые дают информацию о том, с какими словами может использоваться заглавное слово, на основе чего можно выделять устойчивые словосочетания и ассоциативные связи между словами.

Необязательным элементом статьи является значение слова. Оно указывается в скобках после заголовочного слова или после местоименного вопроса в многозначных словах в том случае, если с этим связана форма управляемого слова. Многозначные слова приводятся в одной статье, при этом отдельные значения, если с ними связаны различные МУ нумеруются, омонимы же приводятся в разных статьях.

Так, например, для слова ***подозревать*** словарная статья выглядит следующим образом:

подозревать *кого-л. в ч е м и о ч е м.* ***1.*** *в ч е м (иметь подозрение против кого-л.). Подозревать в обмане. Подозревать в неверности.* ***2.*** *о ч е м (предполагать, догадываться). Дубов и не подозревал о сложных Морозкиных переживаниях (Фадеев).*

Как видно из приведенного примера, структуру каждой словарной статьи образуют элементы, расположенные в четкой последовательности, их можно выделить автоматически. Следовательно, из словаря можно извлекать знания, автоматически заполняя поля записей базы данных МУ.

Методика автоматизированного наполнения базы данных МУ слов русского языка состоит в следующем.

1. Предобработка текста словаря.

2. Автоматическое заполнение полей «написание слова», «морфологическая информация», «толкование слова», «МУ» по словарным статьям предобработанного словаря.

3. Автоматизированное заполнение лингвистом поля «семантический класс» с помощью инструментария работы с базой данных МУ слов русского языка.

Предобработка текста словаря состоит в том, что для каждого многозначного слова статья разбивается на несколько статей, количество которых соответствует числу значений. Так, для приведенного выше примера в результате предобработки получаем 2 статьи следующего содержания:

подозревать *в ч е м (иметь подозрение против кого-л.). Подозревать в обмане. Подозревать в неверности.*

подозревать *о ч е м (предполагать, догадываться). Дубов и не подозревал о сложных Морозкиных переживаниях (Фадеев).*

Для создания процедуры автоматического заполнения полей «написание слова», «морфологическая информация», «толкование слова», «МУ» потребовалось выделить ключевые слова МУ (местоименные вопросы, разделители валентностных гнезд, разделители моделей управления слова), последовательности символов, обозначающие начало или окончание определенного элемента словарной статьи, а также сформировать правила разделения словарной статьи на отдельные поля.

Для заполнения поля «семантический класс» разработан инструментарий, позволяющий также пополнять и редактировать разработанную базу данных МУ слов русского языка. В настоящее время нами используется семантическая классификация предикатов, разработанная на основе семантической классификации Л.Г. Бабенко [6].

Так, для нашего примера слову ***подозревать*** в базе данных МУ будут соответствовать 2 записи, соответствующие приведенным в таблице 1 данным.

Таблица 1 – Примеры МУ слова *подозревать*

Написание слова	МИ	Толкование слова	Семантический класс	МУ
подозревать	Гл. н. вид неперех.	иметь подозрение против кого-л.	Предложения, отображающие ситуацию эмоционально-оценочного отношения	2_1- N2 2_2- N(в)6
подозревать	Гл. н. вид неперех.	предполагать, догадываться	Предложения, отображающие ситуацию воображения и предположения	2- N(о)6

В табл. 1 столбец «МУ» содержит описание заполнения валентно обусловленных ячеек правосторонних актантов.

Этапы предобработки текста словаря и автоматического заполнения полей привязаны к формату и структуре словарных статей выбранного словаря-источника, для каждого словаря-источника на данном этапе необходима разработка отдельных процедур обработки словарных статей. После выполнения этих двух этапов получаем данные в некотором едином представлении, которое может быть использовано системами семантико-синтаксического анализа предложений.

Выводы

В данной работе разработана методика извлечения знаний из лингвистических словарей для наполнения базы данных моделей управления слов русского языка. В процессе проектирования базы данных моделей управления слов русского языка указан способ ее использования системой семантико-синтаксического анализа, разработана структура базы данных. С целью выбора источника для автоматизированного наполнения базы данных моделей управления слов русского языка рассмотрены тексты нескольких словарей. Разработана методика автоматизированного наполнения базы данных моделей управления слов русского языка по тексту выбранного словаря.

Несмотря на обилие синтаксических, семантических словарей, словарей моделей управления, находящихся в электронном виде в открытом доступе, универсального подхода на базе их совместного использования для автоматического наполнения базы данных моделей управления слов русского языка не существует, поскольку способы словарного представления знаний в имеющихся словарях различны. Это приводит к необходимости разработки и распространению стандартов и совместимых лингвистических ресурсов.

Аппарат моделей управления для описания синтаксиса естественного языка позволяет повысить точность синтаксического представления и обеспечить фиксирование стилистических особенностей. В связи с чем представляется перспективным создание программных компонент, поддерживающих предложенную методику автоматического формирования множества моделей управления, а метод описания синтаксиса языка с помощью аппарата моделей управления дает возможность описывать все языковые аспекты (синтаксический, семантический и прагматический) в рамках одной структуры, что позволит существенно увеличить скорость анализа текста и повысить его качество.

Литература

1. Мельчук И.А Опыт теории лингвистических моделей : «Смысл-Текст». Семантика, синтаксис / Мельчук И.А – М. : Школа «Языки русской культуры», 1999. – 992 с.
2. Золотова Г.А. Синтаксический словарь русского языка / Золотова Г.А. – М. : Наука, 1988. – 440 с.
3. Васильев Л.М. Предикаты чувственно-эмоционального переживания и волевых усилий : Системный семантический словарь русского языка / Васильев Л.М. – Уфа : Башкирский государственный университет, 2004. – 310 с.
4. Розенталь Д.Э. Управление в русском языке / Розенталь Д.Э. – М. : Книга, 1986. – 173 с.
5. Ермаков А.Е. Извлечение знаний из текста и их обработка: состояние и перспективы / А.Е. Ермаков // Информационные технологии. – 2009. – № 7.
6. Русские глагольные предложения: Экспериментальный синтаксический словарь / Под общ. ред. Л. Г. Бабенко. – М.: Флинта: Наука, 2002. – 462 с.

Literatura

1. Mel’chuk I.A Opyt teorii lingvisticheskih modelej “Smysl-Tekst”. Semantika, sintaksis. M.: Shkola “Jazyki russkoj kul’tury". 1999. 1992 p.
2. Zolotova G.A. Sintaksicheskij slovar’ russkogo jazyka. M.: Nauka, 1988. 440 p.
3. Vasil’ev L.M. Predikaty chuvstvenno-jemocional'nogo perezhivanija i volevyh usilij : Sistemnyj semanticheskij slovar' russkogo jazyka. Ufa : Bashkirskij gosudarstvennyj universitet. 2004. 310 p.
4. Rozental’ D.Je. Upravlenie v russkom jazyke. M.: Kniga. 1986. 173 p.
5. Ermakov A.E. Informacionnye tehnologii. 2009. №7.
6. Babenko L. G. Russkie glagol’nye predlozhenija: Jeksperimental'nyj sintaksicheskij slovar’. M.: Flinta: Nauka. 2002. 462 p.

RESUME

A.O. Zhuravlyov

Creating of database of governance models for Russian words

A database of governance models for Russian words can help with automatic semantic and syntactic analysis of the text. Establishment of that database manually is a time-consuming process that requires attracting of highly qualified linguists. Thereby, the extraction of knowledge from linguistic dictionaries containing governance model of words is an important problem. Its solution should allow to implement an effective procedure of automated filling the corresponding database.

Despite the abundance of available electronic versions of syntactic, semantic dictionaries and governance models dictionaries, a universal approach based on their common use for an establishment of Russian words governance models database do not exist, because the methods a dictionary knowledge representation in available dictionaries are different.

In this paper we propose a scheme of sentences semantic and syntactic analysis, using the database of Russian words governance, and designed the structure of database governance models. Each governance model is a subcategorization frame, which contains actant's cases and prepositions, that govern of them. In addition, the methods of automated filling of the database is developed. In order to provide possibility of different interpretation of word and partially removing morphological homonymy as source for database filling it is choosed Rosenthal's governance dictionary.

Using of governance models apparatus to describe the syntax of natural language make it possible to improve the accuracy of syntactic representation. Creation of software components that support the proposed method for automatically generating of sets of governance models should allow to substantially increase the text analysis speed and to improve its quality. This suggests prospectivity directions of the selected studies.

Статья поступила в редакцию 02.11.2012.